Integration of Real-World Graph Datasets into EasyGraph

July 2025

1 Introduction

This report documents the integration of eleven well-known graph datasets into the EasyGraph Python library. Each dataset was encapsulated in a dedicated class, with support for download, preprocessing, and loading into an EasyGraph-compatible format. The datasets cover a variety of domains and structures: citation networks, co-purchase graphs, social media interactions, road networks, and web hyperlinks.

Each dataset includes the following:

- A description and primary usage
- Dataset size (nodes, edges, features, classes)
- Directed or undirected nature
- What the dataset is best used for or famous for
- Reference of the original dataset
- Example usage output

2 Datasets

2.1 Amazon Photo

Description: A co-purchase graph of Amazon photo products.
Use Case: Node classification based on product categories.
Famous For: Semi-supervised learning benchmarks (used in PyG, DGL).
Size: 7,650 nodes, 119,082 edges, 745 features, 8 classes.
Directed: No
Dataset: `https://github.com/graphdml-uiuc-jlu/geom-gcn`

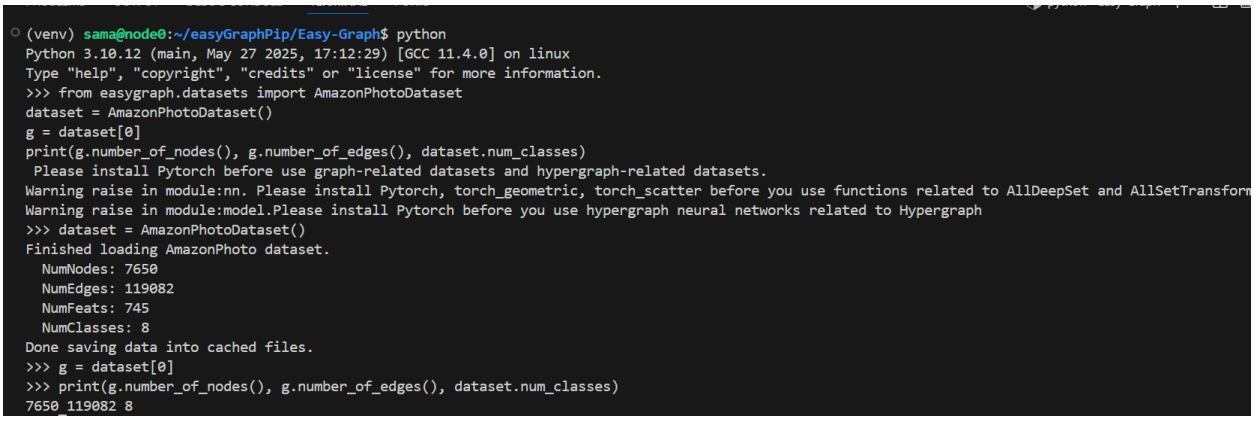

Figure 1: Amazon photos dataset

2.2 Arxiv HEP-TH

Description: Citation network of high-energy physics papers.
Use Case: Link prediction, network evolution.
Famous For: Studying densification laws (Leskovec et al.).
Size: 27,770 nodes, 352,807 edges.
Directed: Yes
Dataset: `https://snap.stanford.edu/data/cit-HepTh.html`

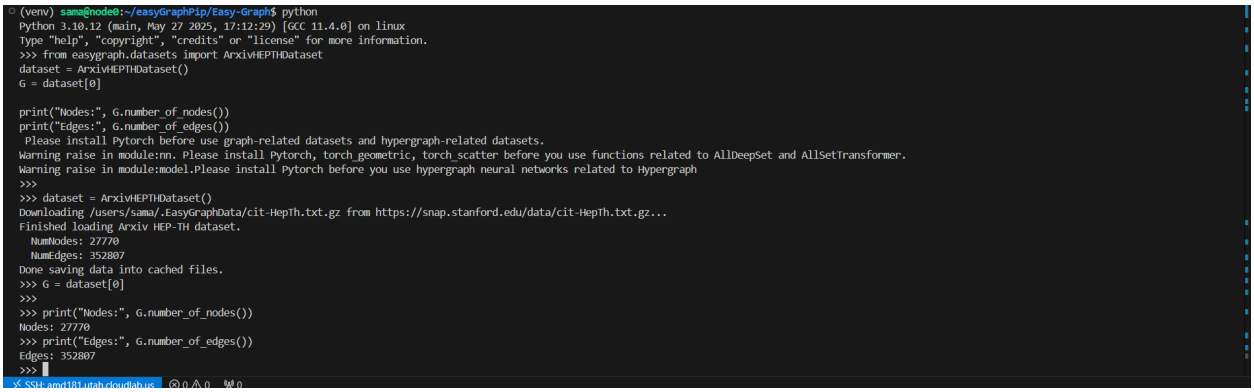

Figure 2: Arxiv HEP-TH dataset

2.3 Coauthor CS

Description: Authors connected by co-authorship in computer science.
Use Case: Node classification and community detection.
Famous For: Used in Cluster-GCN and other scalable GNN benchmarks.
Size: 18,333 nodes, 81,894 edges, 6,805 features, 15 classes.
Directed: No
Dataset: `https://github.com/pyg-team/pytorch_geometric/blob/master/torch_geometric/datasets/coauthor.py`

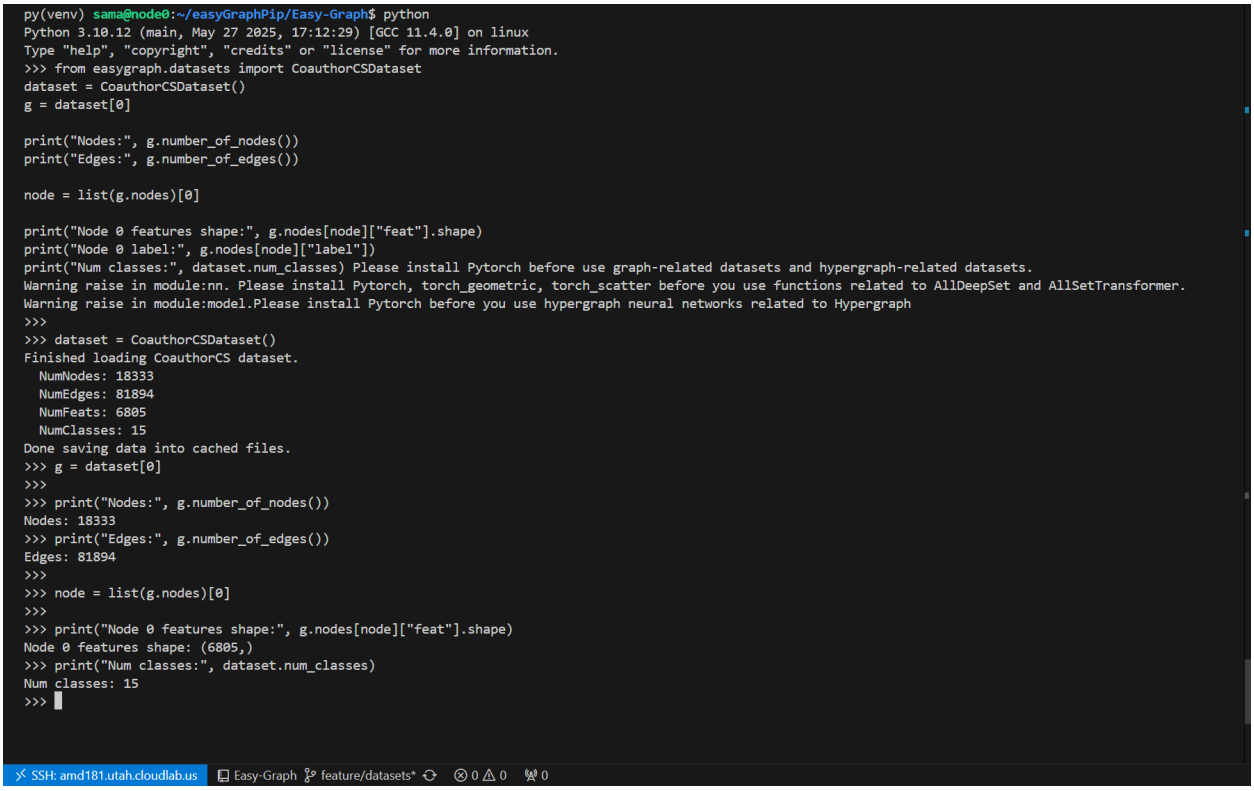

Figure 3: Coauthor CS dataset

2.4 Facebook Ego

Description: Ego-centric friend networks on Facebook.
Use Case: Community detection, link prediction.
Famous For: Social circle prediction (SNAP EgoNets).
Size: 3,959 nodes, 84,243 edges.
Directed: No
Dataset: `https://snap.stanford.edu/data/ego-Facebook.html`

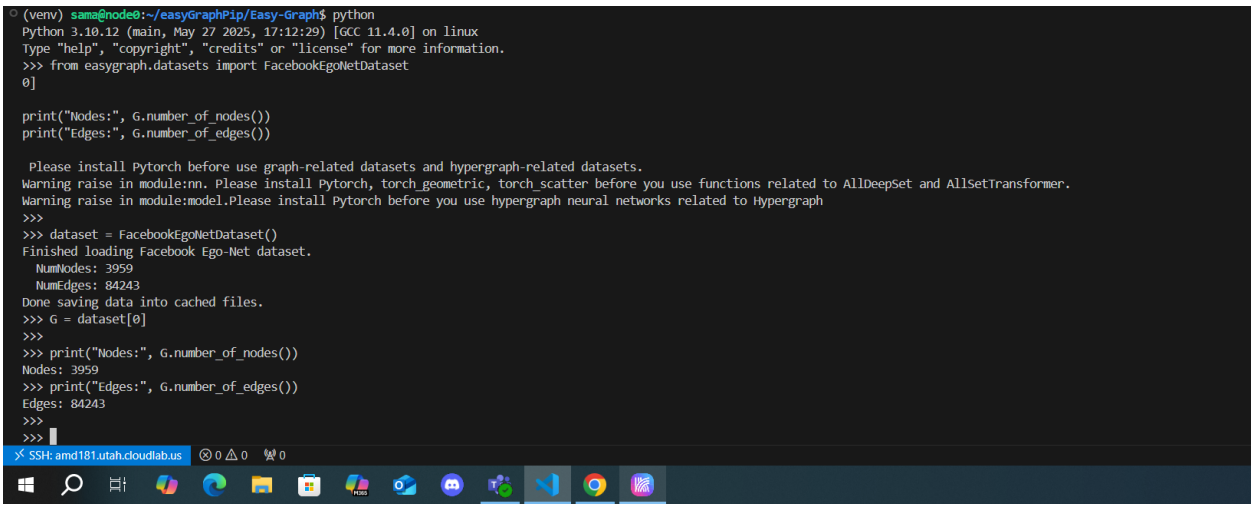

Figure 4: Facebook ego dataset

2.5 Flickr

Description: Images on Flickr connected by common tags.
Use Case: Node classification using image features.
Famous For: GraphSAINT benchmark for sampling-based GNN training.
Size: 89,250 nodes, 449,878 edges, 500 features, 7 classes.
Directed: No
Dataset: `https://github.com/GraphSAINT/GraphSAINT`

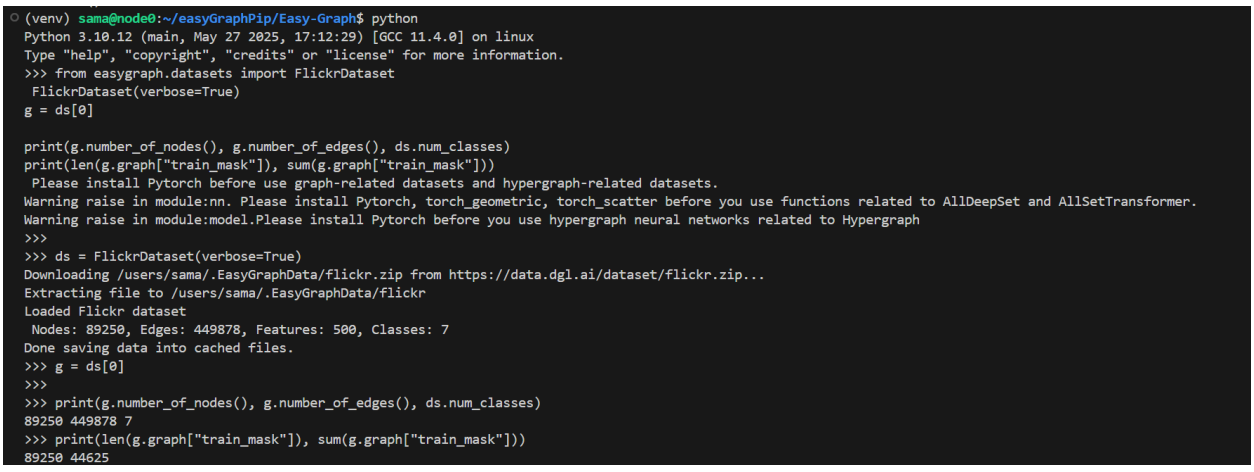

Figure 5: Flickr dataset

2.6 GitHub Users

Description: Social graph of GitHub users (followers).
Use Case: Node classification, influence modeling.
Famous For: Used in GraphSAINT and SNAP social graphs.
Size: 37,700 nodes, 289,003 edges, 5,575 features, 2 classes.
Directed: Yes
Dataset: `https://github.com/williamleif/GraphSAGE`

Figure 6: Github users dataset

2.7 Reddit (Posts)

Description: Reddit posts connected via comment/reply graph.
Use Case: Large-scale node classification by subreddit.
Famous For: Used in inductive GNN training (e.g., GraphSAGE).
Size: 232,965 nodes, 5,754,911 edges, 602 features, 41 classes.
Directed: No
Dataset: `https://github.com/williamleif/GraphSAGE`

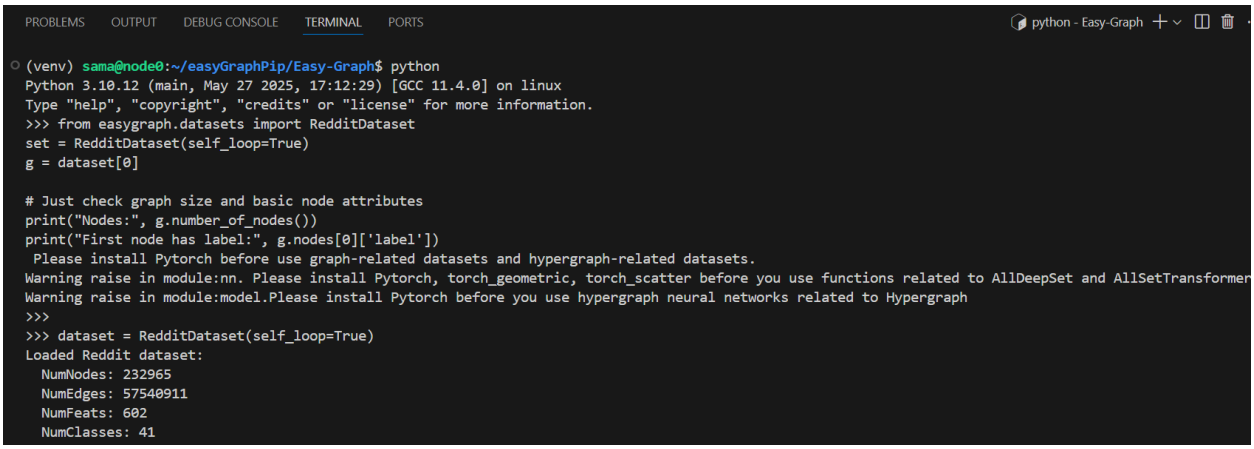

Figure 7: Reddit posts dataset

2.8 RoadNet-CA

Description: Road network of California from intersections.
Use Case: Shortest paths, routing, connected components.
Famous For: Studying spatial networks and road connectivity.
Size: 1,965,206 nodes, 2,766,607 edges.
Directed: No
Dataset: `https://snap.stanford.edu/data/roadNet-CA.html`

Figure 8: RoadNet-CA dataset

2.9 Twitter Ego

Description: Twitter ego network built from user follows.
Use Case: Community analysis, influence tracking.
Famous For: Early SNAP ego-centric dataset.
Size: 81,306 nodes, 1,342,310 edges.
Directed: Yes
Dataset: `https://snap.stanford.edu/data/egonets-Twitter.html`

Figure 9: Twitter ego dataset

2.10 WikiTopCats

Description: Directed graph of Wikipedia articles linked by top-level categories.
Use Case: Multi-label classification, higher-order clustering, community detection.
Famous For: Local clustering benchmarks, overlapping community research (KDD 2017).
Size: 1,791,489 nodes, 28,511,807 edges.
Directed: Yes
Dataset: `https://snap.stanford.edu/data/wiki-topcats.html`

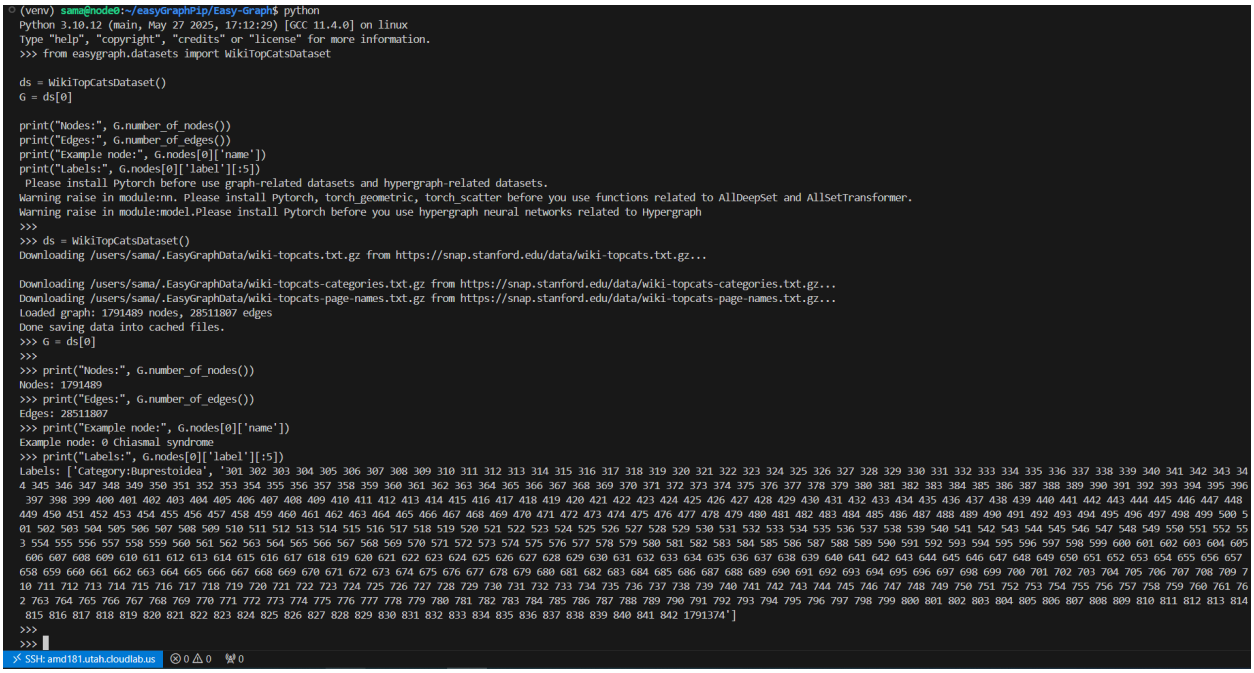

Figure 10: WikiTopCats dataset

2.11 Web-Google

Description: Web pages and hyperlinks from Google's crawl.
Use Case: PageRank, centrality, structural analysis.
Famous For: SNAP web datasets, PageRank studies.
Size: 875,713 nodes, 5,105,039 edges.
Directed: Yes
Dataset: `https://snap.stanford.edu/data/web-Google.html`

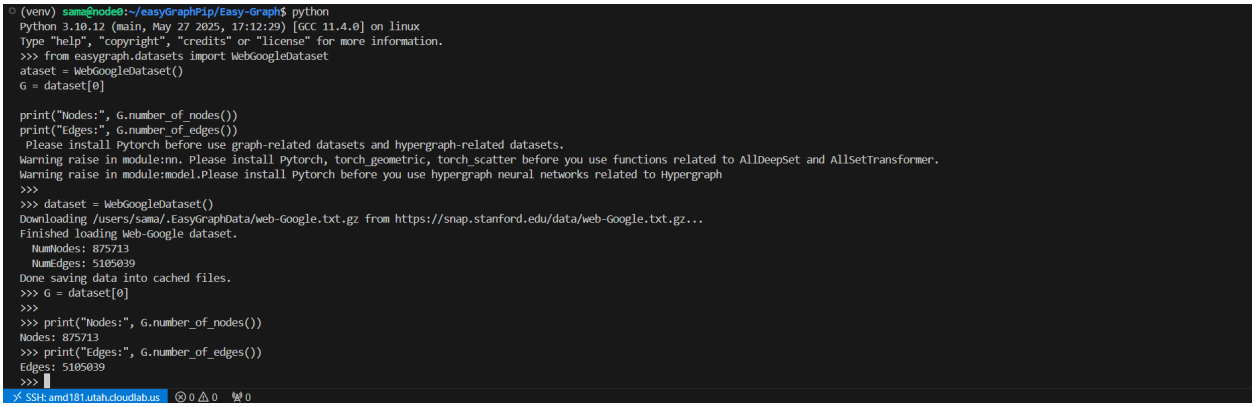

Figure 11: Web-Google dataset

3 Conclusion

The addition of these datasets significantly expands EasyGraph's built-in coverage, enabling rapid experimentation across domains. They support benchmarking of GNNs, topological algorithms, community detection, and more. All datasets follow a standardized interface with support for optional features, labels, and train/val/test masks when applicable.